

日本国特許庁 JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 Date of Application:

2003年 3月 7日

出 願 番 号 Application Number:

人

特願2003-061403

[ST. 10/C]:

[J P 2 0 0 3 - 0 6 1 4 0 3]

出 願
Applicant(s):

株式会社日立製作所

2003年 7月31日

特許庁長官 Commissioner, Japan Patent Office 今井康



【書類名】

特許願

【整理番号】

NT02P0644

【提出日】

平成15年 3月 7日

【あて先】

特許庁長官 殿

【国際特許分類】

G06F 3/06

【発明者】

【住所又は居所】 神奈川県川崎市麻生区王禅寺1099番地 株式会社日

立製作所 システム開発研究所内

【氏名】

田中 勝也

【発明者】

【住所又は居所】

神奈川県川崎市麻生区王禅寺1099番地 株式会社日

立製作所 システム開発研究所内

【氏名】

上村 哲也

【特許出願人】

【識別番号】

000005108

【氏名又は名称】

株式会社日立製作所

【代理人】

【識別番号】

100068504

【弁理士】

【氏名又は名称】 小川 勝男

【電話番号】

03-3661-0071

【選任した代理人】

【識別番号】

100086656

【弁理士】

【氏名又は名称】

田中 恭助

【電話番号】

03-3661-0071

【選任した代理人】

【識別番号】 100094352

【弁理士】

【氏名又は名称】 佐々木 孝

【電話番号】 03-3661-0071

【手数料の表示】

【予納台帳番号】 081423

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【プルーフの要否】 要



【書類名】 明細書

【発明の名称】 ディスクアレイ装置および障害回復制御方法

【特許請求の範囲】

【請求項1】

チャネルパスを介して上位装置に接続されたディスクコントローラと、上記ディスクコントローラに接続された保守端末と、ディスクチャネルを介して上記ディスクコントローラに接続されたディスクアレイとからなり、

上記ディスクアレイが、それぞれ基板内配線に接続された複数のディスクドライブを搭載した複数のドライブ基板と、上記各ドライブ基板の基板内配線をディスクチャネル用の配線と接続するための複数のコネクタを備えた共通基板とからなり、

上記複数のドライブ基板が上記複数のコネクタを介して上記共通基板に着脱可能に搭載され、上記ディスクコントローラが、上記ディスクアレイ内のドライブ基板の一部を予備系、残りを現用系と定義し、現用系のドライブ基板群において互いに異なるドライブ基板上に搭載されたN+1個(N≥2)のディスクドライブを論理グループとして管理し、各論理グループ内で生成された誤り訂正情報の格納領域を上記複数のディスクドライブに分散的または特定ディスクドライブに固定的に割当てて、上記ディスクアレイのデータの書き込みと読み出しを制御するようにしたディスクアレイ装置において、

上記ディスクコントローラが、上記ディスクアレイ内の何れかの現用系ディスクドライブが障害ドライブとなった時、上記障害ドライブが搭載された障害基板上の各ディスクドライブの格納データと同一のデータを上記予備系のドライブ基板群の中から選択された交替基板上のディスクドライブに蓄積した後、上記障害基板上の各ディスクドライブが所属する論理グループを上記交替基板上のディスクドライブを含む新たな構成に再編成するための手段と、

上記論理グループ再編成の完了後に、上記保守端末に対して上記障害基板が交換可能状態となったこと通知するための手段を備えたことを特徴とするディスクアレイ装置。

【請求項2】

2/



前記論理グループの再編成手段が、前記障害ドライブに代わる前記交替基板上の1つの交替ディスクドライブに対して、上記障害ドライブと同一論理グループに所属した他の複数のディスクドライブからの読み出しデータに基づいて再生されたデータを書き込むための手段と、前記障害基板上に搭載された正常ディスクドライブからの読み出しデータを前記交替基板上の別の交替ディスクドライブに順次に複写するための手段とを含むことを特徴とする請求項1に記載のディスクアレイ装置。

【請求項3】

前記共通基板が、前記ディスクチャネル用の配線から前記ドライブ基板接続用の各コネクタを選択的にバイパスするための複数のバイパス回路を備え、

前記ディスクコントローラが、前記論理グループ再編成の完了後に、前記障害 基板の接続コネクタと対応するバイパス回路をバイパス状態に切替えるための手 段を備えたことを特徴とする請求項1または請求項2に記載のディスクアレイ装 置。

【請求項4】

前記ディスクコントローラが、前記障害基板の接続コネクタに復旧基板が再接 続された時、該復旧基板を予備系のドライブ基板として管理することを特徴とす る請求項1または請求項2に記載のディスクアレイ装置。

【請求項5】

前記ディスクコントローラが、前記ディスクアレイを構成するドライブ基板を 正常状態、交換待ち状態、予備状態の順に遷移する状態コードによって管理する ための基板管理テーブルを備え、状態コードが正常状態の基板を前記現用系、予 備状態の基板を前記予備系とすることを特徴とする請求項1または請求項2に記 載のディスクアレイ装置。

【請求項6】

前記共通基板上で前記ドライブ基板接続用のコネクタがX、Y軸上の座標値で特定される2次元配置を有し、各ドライブ基板上の複数のディスクドライブがZ軸方向に配列され、

前記ディスクコントローラが、互いに同一のX座標値、Z座標値を有し、Y座

標値が異なるN+1個のディスクドライブによって前記論理グループを形成し、

前記論理グループ再編成手段が、前記予備系のドライブ基板群の中から前記障 害基板と同一のY座標値をもつドライブ基板を前記交替基板として選択し、上記 故障基板上の複数のディスクドライブと上記交替基板上の交替ディスクドライブ とをそれぞれのZ座標値に従って対応付けることを特徴とする請求項1または請 求項2に記載のディスクアレイ装置。

【請求項7】

上位装置と保守端末に接続されたディスクコントローラと、ディスクチャネル を介して上記ディスクコントローラに接続されたディスクアレイとからなり、

上記ディスクアレイが、それぞれ基板内配線に接続された複数のディスクドライブを搭載した複数のドライブ基板と、上記各ドライブ基板の基板内配線をディスクチャネル用の配線と接続するための複数のコネクタを備えた共通基板とからなり、

上記複数のドライブ基板が上記複数のコネクタを介して上記共通基板に着脱可能に搭載され、上記ディスクコントローラが、上記ディスクアレイ内のドライブ基板の一部を予備系、残りを現用系と定義し、現用系のドライブ基板群において互いに異なるドライブ基板上に搭載されたN+1個(N≥2)のディスクドライブを論理グループとして管理し、各論理グループ内で生成された誤り訂正情報の格納領域を上記複数のディスクドライブに分散的または特定ディスクドライブに固定的に割当てて、上記ディスクアレイのデータの書き込みと読み出しを制御するようにしたディスクアレイ装置における上記ディスクコントローラによる障害回復制御方法であって、

上記ディスクアレイ内の何れかの現用系ディスクドライブが障害ドライブとなった時、上記予備系のドライブ基板群の中から、上記障害ドライブが搭載された 障害基板に代わる交替基板を選択するステップと、

上記障害ドライブが搭載された障害基板上の各ディスクドライブの格納データ と同一のデータを上記予備系のドライブ基板群の中から選択された交替基板上の ディスクドライブに蓄積するステップと、

上記障害基板上の各ディスクドライブが所属する論理グループを上記交替基板

上のディスクドライブを含む新たな構成に再編成するステップと、

上記論理グループの再編成後に、上記保守端末に対して上記障害基板が交換可能状態となったこと通知するステップとを含むことを特徴とする障害回復制御方法。

【請求項8】

前記交替基板上のディスクドライブへのデータ蓄積ステップが、

前記障害ドライブに代わる前記交替基板上の1つの交替ディスクドライブに対して、上記障害ドライブと同一論理グループに所属した他の複数のディスクドライブからの読み出しデータに基づいて再生されたデータを書き込むステップと、

前記障害基板上に搭載された正常ディスクドライブからの読み出しデータを前記交替基板上の別の交替ディスクドライブに順次に複写するステップとからなることを特徴とする請求項7に記載の障害回復制御方法。

【請求項9】

前記共通基板が、前記ディスクチャネル用の配線から前記ドライブ基板接続用 の各コネクタを選択的にバイパスするための複数のバイパス回路を備え、

前記ディスクコントローラが、前記論理グループの再編成後に、前記障害基板の接続コネクタと対応するバイパス回路をバイパス状態に切替えるステップを含むことを特徴とする請求項7または請求項8に記載の障害回復制御方法。

【発明の詳細な説明】

[0001]

【発明の属する技術分野】

本発明は、ディスクアレイ装置および障害回復制御方法に関し、更に詳しくは、冗長ディスクドライブの含む複数のディスクドライブで1つの論理グループを 形成してディスク障害に備えるRAID形式のディスクアレイ装置および障害回 復制御方法に関する。

[0002]

【従来の技術】

高性能のコンピュータシステムは、大容量の2次記憶装置を備え、CPU等の上位装置が必要とするデータを上記2次記憶装置に随時にリード・ライトしてい

る。 2 次記憶装置としては、例えば、磁気ディスクや光ディスクのように、ランダムアクセス可能な不揮発性記憶媒体をもつディスク装置が一般的であり、最近では、記憶容量を増加するために多数の小型ディスクドライブ(以下、単にドライブと言う)からなるディスクアレイ装置が主流になってきている。

[0003]

ディスクアレイ装置では、少なくとも1個の冗長ドライブを含む複数のドライブで1つの論理グループを形成し、論理グループ毎にドライブ障害に対処できるようにした RAID (Redundant Array of Inexpensive Disk) 方式が採用されている。

[0004]

RAIDには標準化された幾つかのレベルがあり、例えば、RAID1 (レベル1)のディスクアレイ装置では、データ格納用の各ドライブと対をなして冗長ドライブ (予備ドライブ)を用意しておき、同一データを2つのドライブに並列的に書き込むことによって、何れかのデータドライブに障害が発生したとき、これと対をなす予備ドライブから必要なデータを読み出せるようにしている。

$[0\ 0\ 0\ 5]$

RAID3(レベル3)では、N+1個(N \geq 2)のドライブで1つの論理グループ(パリティグループまたはRAIDグループと言う場合もある)を形成し、そのうちの1個を誤り訂正情報(以下、パリティで代表させる)格納用の冗長ドライブ、残りをデータ格納用ドライブとしている。本明細書では、誤り訂正情報をパリティで代表させるが、各論理グループで生成される誤り訂正情報にはパリティ以外の他の情報を適用できること明らかである。

[0006]

RAID3のディスクアレイ装置では、上位装置からデータブロックの書き込み要求があった時、書き込み要求に付随するデータブロックを固定長(例えば、1バイト長)の複数のサブデータブロックに分割し、これらのサブデータブロックを上記N個のデータ格納用ドライブに順次に振り分けて格納する。冗長ドライブには、データ格納用ドライブ内で互いに同一のアドレスをもつ同一論理グループ内のN個のサブデータブロックから生成された誤り訂正情報(例えば、パリテ

ィデータ)が格納される。上位装置からデータブロックの読み出し要求があった時は、上記N個のデータ格納用ドライブから並列的にサブデータブロックを読み出し、これらのサブデータブロックを所定の順序で結合することによって、元のデータブロックを復元している。

[0007]

RAID4(レベル4)でも、N+1個(N \geq 2)のドライブで1つの論理グループを形成し、そのうちの1個を誤り訂正情報格納用の冗長ドライブ、残りをデータ格納用ドライブとしている。但し、RAID4のディスクアレイ装置では、上位装置からデータブロックの書き込み要求があった時、1つの書き込み要求に付随するデータブロックを上記何れかのデータ格納用ドライブに格納し、次の書き込み要求に付随するデータブロックは別のデータ格納用ドライブに格納する形式で、データの書き込みが行われる。従って、冗長ドライブには、データ格納用ドライブ内で同一のアドレスをもつそれぞれが別のデータブロックに所属するデータ部分から生成された誤り訂正情報が格納されることになる。

[0008]

RAID5(レベル5)では、レベル4と同様に、1つの書き込み要求に付随するデータブロックでのデータ書き込みが行なわれる。但し、誤り訂正情報の格納領域は、レベル4にように特定のディスクドライブに固定的に割当てられるのではなく、論理グループを形成する複数 (N+1) のディスクドライブに分散的に割当てられる。

[0009]

RAID3~RAID5のディスクアレイ装置では、何れかのドライブに障害が発生した場合、障害ドライブと同一論理グループに所属する他のドライブから読み出したデータに基づいて、上記障害ドライブが保持していたデータまたは誤り訂正情報(例えば、パリティデータ)を再生することが可能となる。

[0010]

上述したディスクアレイ装置で、記憶容量を大容量化し、筐体サイズを小型化するためには、狭い空間にできるだけ多数のドライブを実装する必要がある。一般的には、それぞれが制御用のLSIチップを搭載した複数の制御基板と、それ

ぞれが複数のドライブを搭載した多数のドライブ基板とをマザーボード上に並列的に配設されたコネクタに差込み、各ドライブをドライブ基板上の配線を介してマザーボード上のディスクチャネル配線に結合するようにした装置構成が採用される。

[0011]

この構成において、ドライブ障害が発生した時、ドライブ基板から障害ドライブだけを取り外して新たなドライブと交換しようとすると、隣接するドライブ基板の間にドライブ交換作業に必要な空間を残す必要があり、マザーボード上でのドライブ基板の実装密度が低下する。

[0012]

この問題点に着目した従来技術として、例えば、特開平7-230362号公報 (特許文献1)には、マザーボードに多数のドライブ基板を高密度で実装しておき、ドライブ障害が発生した時は、障害ドライブが搭載されているドライブ基板 (障害基板) そのものをマザーボードから取り外し、装置筐体の外部において障害ドライブを正常ドライブに交換し、部品交換されたドライブ基板をマザーボードに再接続するようにしたディスクアレイ装置が提案されている。

$[0\ 0\ 1\ 3\]$

上記構成によれば、部品交換期間中に、障害ドライブのみならず、障害基板上 に搭載されている複数の正常ドライブもディスクアレイから除外されてしまう。

そこで、上記特許文献1では、それぞれ異なるドライブ基板上に搭載されているN+1個のドライブによって各論理グループを形成しておき、上位装置から障害ドライブの格納データ、または障害基板の取り外しによって不在となったドライブの格納データに対して読み出し要求があった時は、障害ドライブまたは不在ドライブと同一論理グループに所属する他の複数のドライブからデータを読み出し、これらのデータによって要求データを再生するようにしている。

[0014]

また、上記特許文献1では、障害基板の取り外し前に障害ドライブに対して書き込み要求があったデータと、障害基板の取り外し後に不在となったドライブ(障害ドライブと複数の正常ドライブ)に対して書き込み要求があったデータにつ いては、ディスク制御装置が備えるキャッシュメモリに格納しておき、部品交換されたドライブ基板がマザーボードに再接続された時、キャッシュメモリから復旧基板上の該当ドライブにデータを書き移すことを提案している。尚、障害ドライブの除去によって読み出し不能となった損失データについては、ドライブ基板が再接続された時、障害ドライブと同一論理グループに所属する他の複数のドライブからの読み出しデータによって再生し、復旧基板上の交替ドライブに書き込むようにしている。

[0015]

上記特許文献1では、更に、各論理グループと対応して予備ドライブを用意しておき、上位装置から不在ドライブに対して書き込み要求のあったデータをキャッシュメモリの代わりに上記予備ドライブに一時的に格納しておき、部品交換されたドライブ基板がマザーボードに再接続された時、上記予備ドライブから復旧基板上の該当ドライブにデータを書き移すことも提案している。また、障害ドライブに代わる予備ドライブの蓄積データが大量になった場合は、部品交換されたドライブ基板が再接続された時、予備ドライブ(または障害ドライブ)と同一論理グループに所属する他の複数のドライブからの読み出しデータによって障害ドライブの損失データを再生し、これらのデータを上記予備ドライブに書き込むことによって、予備ドライブを正常ドライブとして引き続き利用してもよい旨の提案もしている。

[0016]

【特許文献1】

特開平7-230362号

[0017]

【発明が解決しようとする課題】

然るに、特許文献1で提案されたディスクアレイ装置は、筐体サイズの小型化と記憶容量の大容量化に好適なハードウエア構成と言えるが、障害ボードの取り外し期間中に上位装置からデータの読み出し要求があった時、障害ドライブのみならず、不在となった複数の正常ドライブについても、論理グループを利用したデータの再生動作が必要となり、障害基板の取り外し期間中のデータ読み出し要

求に対して応答遅れが発生するという問題があった。

[0018]

また、上記従来技術では、部品交換されたドライブ基板がマザーボードに復帰した時に、キャッシュメモリまたは予備ドライブから復旧ボード上のドライブへのデータ書き込みと、障害ドライブ損失データの再生動作とを実行しているため、ディスクアレイの正常状態への復帰が遅れるという問題があった。

[0019]

本発明の目的は、筐体サイズの小型化と記憶容量の大容量化に適し、ドライブ 障害が発生しても、速やかに正常状態に復帰できるディスクアレイ装置および障 害回復制御方法を提供することにある。

本発明の他の目的は、ドライブ障害が発生した時、保守員による障害ドライブ の部品交換に十分な作業時間を許容できるディスクアレイ装置および障害回復制 御方法を提供することにある。

[0020]

【課題を解決するための手段】

上記目的を達成するため、本発明のディスクアレイ装置および障害回復制御方法では、ドライブ障害が発生した時、障害ドライブを搭載した障害基板上の複数のディスクドライブについて、交替ディスクドライブへのデータの移し替えを行い、論理グループの再編成が完了した後に、障害基板の交換を行うようにしたことを特徴とする。

[0021]

障害ドライブの格納データは、同一論理グループに所属した他のドライブからの読み出しデータに基づいて再生する必要があるが、障害基板上の正常ドライブの格納データは、各ドライブからの読み出しデータをそのまま交替ディスクドライブに複写すれば済むため、正常ドライブから交替ディスクドライブへのデータの移し替えは短時間で終了できる。本発明によれば、論理グループの再編成が完了した後に障害基板の交換を行うようにしているため、障害基板の取り外し期間中でも、通常のデータ・リード/ライトが可能となり、従来技術のように、不在ドライブに対する論理グループによるデータ再生を行う必要がない。

[0022]

更に詳述すると、本発明は、チャネルパスを介して上位装置に接続されたディスクコントローラと、上記ディスクコントローラに接続された保守端末と、ディスクチャネルを介して上記ディスクコントローラに接続されたディスクアレイとからなり、上記ディスクアレイが、それぞれ基板内配線に接続された複数のディスクドライブを搭載した複数のドライブ基板と、上記各ドライブ基板の基板内配線をディスクチャネル用の配線と接続するための複数のコネクタを備えた共通基板とからなり、上記複数のドライブ基板が上記複数のコネクタを介して上記共通基板に着脱可能に搭載され、上記ディスクコントローラが、上記ディスクアレイ内のドライブ基板の一部を予備系、残りを現用系と定義し、現用系のドライブ基板群において互いに異なるドライブ基板上に搭載されたN+1個(N≥2)のディスクドライブを論理グループとして管理し、各論理グループ内で生成された誤り訂正情報の格納領域を上記複数のディスクドライブに分散的または特定ディスクドライブに固定的に割当てて、上記ディスクアレイのデータの書き込みと読み出しを制御するようにしたディスクアレイ装置において、

上記ディスクコントローラが、上記ディスクアレイ内の何れかの現用系ディスクドライブが障害ドライブとなった時、上記障害ドライブが搭載された障害基板上の各ディスクドライブの格納データと同一のデータを上記予備系のドライブ基板群の中から選択された交替基板上のディスクドライブに蓄積した後、上記障害基板上の各ディスクドライブが所属する論理グループを上記交替基板上のディスクドライブを含む新たな構成に再編成するための手段と、

上記論理グループ再編成の完了後に、上記保守端末に対して上記障害基板が交換可能状態となったこと通知するための手段を備えたことを特徴とする。

[0023]

本発明の1 実施例によれば、上記論理グループの再編成手段が、障害ドライブ に代わる交替基板上の1つの交替ディスクドライブに対して、上記障害ドライブ と同一論理グループに所属した他の複数のディスクドライブからの読み出しデー タに基づいて再生されたデータを書き込むための手段と、障害基板上に搭載され た正常ディスクドライブからの読み出しデータを上記交替基板上の別の交替ディ スクドライブに順次に複写するための手段とを含む。

[0024]

本発明の他の特徴は、共通基板が、ディスクチャネル用の配線からドライブ基板接続用の各コネクタを選択的にバイパスするための複数のバイパス回路を備え、上記ディスクコントローラが、上記論理グループ再編成の完了後に、障害基板の接続コネクタと対応するバイパス回路をバイパス状態に切替えるための手段を備えることにある。

[0025]

本発明の更に他の特徴は、上記ディスクコントローラが、ディスクアレイを構成するドライブ基板を正常状態、交換待ち状態、予備状態の順に遷移する状態コードによって管理するための基板管理テーブルを備え、状態コードが正常状態の基板を現用系、予備状態の基板を予備系として扱うことにある。

[0026]

本発明の1実施例によれば、共通基板上で前記ドライブ基板接続用のコネクタがX、Y軸上の座標値で特定される2次元配置を有し、各ドライブ基板上の複数のディスクドライブが2軸方向に配列され、上記ディスクコントローラが、互いに同一のX座標値、Z座標値を有し、Y座標値が異なるN+1個のディスクドライブによって論理グループを形成する。また、前記論理グループ再編成手段が、予備系のドライブ基板群の中から故障基板と同一のY座標値をもつドライブ基板を交替基板として選択し、上記障害基板上の複数のディスクドライブと上記交替基板上の交替ディスクドライブとをそれぞれの2座標値に従って対応付ける。

[0027]

本発明によるディスクアレイ装置における障害回復制御方法は、ディスクコントローラが、

ディスクアレイ内の何れかの現用系ディスクドライブが障害ドライブとなった時、予備系のドライブ基板群の中から、上記障害ドライブが搭載された障害基板に代わる交替基板を選択するステップと、

上記障害ドライブが搭載された障害基板上の各ディスクドライブの格納データ と同一のデータを上記予備系のドライブ基板群の中から選択された交替基板上の ディスクドライブに蓄積するステップと、

上記障害基板上の各ディスクドライブが所属する論理グループを上記交替基板 上のディスクドライブを含む新たな構成に再編成するステップと、

上記論理グループの再編成後に、上記保守端末に対して上記障害基板が交換可能状態となったこと通知するステップとを含むことを特徴とする。

本発明のその他の特徴は、図面を参照した以下の実施例の説明から明らかになる。

[0028]

【発明の実施の形態】

以下、本発明の実施例について、図面を参照して説明する。

図1は、本発明が適用されるディスクアレイ装置の1実施例を示す。

ここに示したディスクアレイ装置は、複数のディスクドライブからなるディスクアレイ5を2台のディスクコントローラ1A、1Bからアクセスする冗長構成となっている。これらのディスクコントローラ1A、1Bは、表示装置9を備えたサービスプロセッサ(SVP)2に接続されている。

[0029]

ディスクコントローラ1Aは、複数本のチャネルパス3A(30A~33A)を介して上位装置となるCPU(図示せず)に接続されるチャネルアダプタ10と、後述するバイパス制御用の信号線を含む複数本のディスクチャネル4A(40A~43A)を介してディスクアレイ5に接続されるディスクアダプタ(DKA)20と、これらの要素を相互に接続する相互結合網40とからなる。

[0030]

本実施例では、各論理グループ(以下、RAIDグループと言う)を構成するディスクドライブ数を4個とし、各ディスクドライブのロジカルユニット数を1個として説明するが、本発明は、RAIDグループのコンポーネントとなるドライブ個数と各ディスクドライブのロジカルユニット数を実施例に限定するものではない。

[0031]

ディスクアレイ5は、複数のディスクドライブPDEVO~PDEV127からなり、これ

らのディスクドライブは、ディスクチャネルと対応した4つのドライブ群(第 0 群:PDEV0~PDEV31、第 1 群:PDEV32~PDEV63、第 2 群:PDEV64~PDEV95、第 3 群:PDEV96~PDEV127)に分割され、各ドライブ群がそれぞれと対応するディスクチャネル 4 0 A ~ 4 3 A、 4 0 B ~ 4 3 Bに接続されている。ディスクコントローラ 1 B も、上記ディスクコントローラ 1 A と同様の構成となっており、ディスクコントローラ 1 B のディスクアダプタ 2 0 は、ディスクチャネル 4 B(4 0 B ~ 4 3 B)を介してディスクアレイ 5 に接続されている。

[0032]

キャッシュメモリ30は、ディスクコントローラ1A、1B内の相互結合網40と接続され、ディスクコントローラ1A、1Bの双方からアクセス可能となっている。

[0033]

図2は、チャネルアダプタ10の構成を示す。

チャネルアダプタ10は、チャネルパス3Aに接続されたホストチャネル・インタフェース11と、相互結合網40に接続されたキャッシュメモリ・インタフェース12と、SVP2に接続するためのネットワークインタフェース13と、CPUとの間でのデータ転送を制御するためのプロセッサ14と、該プロセッサが参照する各種のテーブルや実行すべきソフトウェアを格納したローカルメモリ15と、これらの要素間の相互接続するプロセッサ周辺制御部16からなる。

[0034]

ホストチャネル・インタフェース11は、チャネルパス3A上の制御プロトコルとディスクコントローラ内部の制御プロトコルとの間の変換機能を有し、ホストチャネル・インタフェース11とキャッシュメモリ・インタフェース12との間は信号線17によって接続されている。

[0035]

図3は、ディスクアダプタ20の構成を示す。

ディスクアダプタ20は、相互結合網40に接続されたキャッシュメモリ・インタフェース21と、ディスクチャネル4Aに接続されたディスクチャネル・インタフェース22と、SVP2に接続するためのネットワークインタフェース2

3と、プロセッサ24と、該プロセッサが参照する各種のテーブルや実行すべき ソフトウェアを格納したローカルメモリ25と、これらの要素間の相互接続する プロセッサ周辺制御部26とからなる。

[0036]

キャッシュメモリ・インタフェース21とディスクチャネル・インタフェース22との間は信号線27によって接続されている。ディスクチャネル・インタフェース22は、ディスクコントローラ内部の制御プロトコルと、ディスクチャネル4A上の制御プロトコル、例えば、SCSIとの間の変換機能を備えている。

[0037]

図4は、本発明によるディスクアレイ5の実装構造の1例を示す。

ディスクアレイ5は、マザーボード(共通基板)6上のコネクタ群8に着脱自在に搭載された複数のドライブ基板BRD(BRD0~BRD31)からなり、本実施例では、各ドライブ基板BRDに4つのディスクドライブPDEV(i)~PDEV(i+3)が搭載されている。上記マザーボード6には、実際には、上記ドライブ基板の他に、例えば、チャネルアダプタ10やディスクアダプタ20を構成する各種のLSIチップが搭載された制御基板や、キャッシュメモリ30用のメモリLSIを搭載したメモリ基板など、他の複数の基板が搭載されるが、ここでは、ドライブ基板のみに着目して説明する。

[0038]

マザーボードの左上を基点としてディスクアレイ 5 にX、Y、Z 軸を設定し、X 軸に沿って配列されたドライブ基板に着目すると、最上段(Y=0)のドライブ基板BRD0~BRD7に搭載されたディスクドライブPDEV0~PDEV31は、図 1 に示した第 0 群のディスクドライブに相当している。これらの第 0 群のディスクドライブは、ポートバイパス回路(PBC) 7-0 A~ 7-7 A、7-0 B~ 7-7 B を介して、マザーボード 6 上の第 0 ディスクチャネル 4 0 A、4 0 Bに接続されている。

[0039]

一同様に、第2段(Y=1)のドライブ基板BRD8~BRD15に搭載された第1群の ディスクドライブPDEV32~PDEV63は、マザーボード6上の第1ディスクチャネル 4 1 A、4 1 Bに接続され、第 3 段(Y = 2)のドライブ基板BRD16~BRD23に搭載された第 2 群のディスクドライブPDEV64~PDEV95は、マザーボード 6 上の第 2 ディスクチャネル 4 2 A、 4 2 Bに、第 4 段(Y = 3)のドライブ基板BRD24~BRD31に搭載された第 3 群のディスクドライブPDEV96~PDEV127は、マザーボード6 上の第 3 ディスクチャネル 4 3 A、 4 3 Bに接続されている。

[0040]

上記ディスクアレイ 5 において、左端で Y 軸に沿って配列された 4 枚のドライブ基板BRD0、BRD8、BRD16、BRD24に破線で示したように、互いに同一の X 座標値、 Z 座標値をもつ縦方向に配列された 4 つのディスクドライブによって、R A I Dグループが形成される。本実施例では、ドライブ基板BRD0、BRD8、BRD16、BRD 24上のディスクドライブで形成される R A I Dグループを Z 座標順に VDE V0~VDE V3と呼ぶ。以下、 X 座標の方向に Z 座標順に、 R A I Dグループを VDE V4、 VDE V5、 VDE V6、 VDE V7、 ··· VDE V31と呼び、これらの R A I Dグループ名に付された数字 0、1、2、 ··· を R A I Dグループ識別子と定義する。

[0041]

各RAIDグループ内では、例えば、各ディスクドライブのメモリ空間を複数のストライピング領域に分割し、ストライピング領域によって誤り訂正情報(以下、パリティで代表させる)格納用のディスクドライブが異なるように、パリティ用ドライブの分散化が図られている。

[0042]

実際の応用においては、ディスクアレイ5に含まれる全てのディスクドライブがRAIDグループに組み込まれる訳ではなく、そのうちの何個かは予備ドライブとして待機状態となる。以下の説明では、ディスクアレイ5の右端に位置した縦2列のドライブ基板BRD6~BRD30とBRD7~BRD31を予備基板とする。現用系の何れかのドライブ基板でディスクドライブに障害が発生した時、上記予備基板のうちの1つが障害基板に代わる交替基板に選択され、交替基板上のディスクドライブが現用系のRAIDグループに組み込まれる。

[0043]

図5と図6は、本発明のディスクアレイ5において、現用系ディスクドライブ

のうちの1つ、例えば、ディスク基板BRDO上の左から2番目のディスクドライブ に障害が発生した場合のRAIDグループ再編成方法を説明するための図である

[0044]

本発明のディスクアレイ5では、同一ディスク基板上に搭載された複数のディスクドライブが、それぞれ別々のRAIDグループに所属しているため、何れかのディスクドライブに障害が発生した時、障害ドライブを含むディスク基板をマザーボードから外す前に、同一ディスク基板上に搭載された正常なディスクドライブについても、RAIDグループを再編成しておく。

[0045]

図5において、ディスク基板BRDO上でドライブ番号「1」をもつディスクドライブが障害ドライブPDEV(E)となった時、PDEV(E)が所属するRAIDグループVDEVIが異常RAIDグループとなる。この場合、ディスク基板BRDO上に搭載されたドライブ番号「0」、「2」、「3」をもつ正常ディスクドライブPDEV(EZ1)、PDEV(EZ2)、PDEV(EZ3)は、障害基板BRDOをマザーボード6から抜き取った時、それぞれが所属する正常なRAIDグループVDEVO、VDEV2、VDEV3から除外されてしまうため、これらの正常ディスクドライブPDEV(EZ1)、PDEV(EZ2)、PDEV(EZ3)についてもRAIDグループの再編成が必要となる。本明細書では、障害基板の抜き取りに伴ってRAIDグループ再編成の対象となる正常ディスクドライブを連累ドライブと呼ぶことにする。

[0046]

本発明の特徴は、障害回復処理において、障害ドライブPDEV(E)のみならず、連累ドライブPDEV(EZ1)、PDEV(EZ2)、PDEV(EZ3)についても、RAIDグループの再編成を行うことにある。このため、予備基板群BRD6~BRD30、BRD7~BRD31の中から、例えば、障害基板BRD0と同一のY座標値をもつ予備基板BRD7を交替基板として選択し、この交替基板上の各ディスクドライブに、上記障害基板上の各ドライブの格納データと同一のデータを移し替える。

[0047]

本実施例の場合、障害回復処理の結果、例えば、図6に示すように、RAID

グループVDEV0~VDEV3が交替基板上のディスクドライブPDEV(RZ0)~PDEV(RZ3)を含む形に再編成される。ここで、障害ドライブPDEV(E)に代わる交替ドライブPDE V(RZ1)に対しては、RAIDグループVDEV1を構成している他の正常ドライブ、ここではドライブ番号「33」、「65」、「97」をもつディスクドライブから読み出されたデータに基づいて、障害ドライブPDEV(E)で失ったデータを回復した後、順次にデータ書込みが行われる。一方、連累ドライブに代わる交替ドライブPDEV(RZ0)、PDEV(RZ2)、PDEV(RZ3)に対しては、連累ドライブを正常にアクセスできるため、それぞれ対応する連累ドライブから読み出したデータを順次に複写すればよい。

[0048]

図7は、本発明のディスクアレイ5におけるディスクドライブの状態遷移を示す。

現用系の同一ドライブ基板上に搭載された 4 個のディスクドライブPDEV (EZ0) ~ PDEV (EZ3) は、最初、正常状態 S T 0 で動作している。何れかのドライブ、例えば、ディスクドライブPDEV (EZ0) に異常(E V T 1)が発生すると、障害ドライブPDEV (EZ0) の状態コードは異常状態 S T 2 に遷移する。

[0049]

障害基板に代わる交替基板が選択され、RAIDグループ(VDEV)の再編成(EVT2)が完了すると、障害基板上の各ディスクドライブの状態は交換待ち状態ST3に遷移する。障害基板がマザーボードから除去され、部品交換した後、マザーボードに正常な基板が挿入(EVT3)されると、復旧基板上の各ディスクドライブPDEV(EZ0)~PDEV(EZ3)は、予備状態ST1となる。これらのディスクドライブは、その後に他のディスクドライブに異常が発生し、RAIDグループ(VDEV)の再編成(EVT4)で現用系ドライブ群に加えられた時、正常状態ST0に遷移する。

[0050]

図8は、ディスク基板上のディスクドライブ群をディスクチャネルに接続する ためのポートバイパス回路7の構成を示す。

各ディスクチャネルは、複数のディスクドライブをシリアルに接続したアクセ

スループを形成している。ポートバイパス回路 7 は、ディスクドライブ(またはドライブ基板)の接続ポートとして機能する回路であって、ディスクチャネル(アクセスループ)の入回線 4 0 A からの信号をディスクドライブの入力端 I N に供給するためのドライバ回路 7 1 と、上記入回線 4 0 A からの入力信号とディスクドライブの出力端 O U T からの信号の何れか一方を選択して、ディスクチャネル(アクセスループ)の出回線 4 0 A'に出力するマルチプレクサ 7 2 からなっている。

マルチプレクサ72は、例えば、選択信号SELECTが"1"状態の時は、ディスクドライブの出力信号を選択し、選択信号が"0"状態の時は、ディスクチャネルの入回線40Aからの入力信号を選択し、出回線40Aを介して次のポートに供給する。

[0051]

図9は、第0群のディスクドライブPDEVO〜PDEV31と、第0ディスクチャネル40A、40Bとの接続関係と、バイパス制御信号線SEL-OA〜SEL-7A、SEL-0B〜SEL-7Bを示す。

マザーボード 6 上に配線された第 0 ディスクチャネル 4 0 A、 4 0 B は、第 0 群のドライブ基板BRD0~BRD7との接続コネクタの近傍に位置して、それぞれポートバイパス回路 7-0 A~ 7-7 A、 7-0 B~ 7-7 Bを備えており、これらのポートバイパス回路を介して、ドライブ基板BRD0~BRD7上のアクセスループ 4 0 A -0 ~ 4 0 A -7 、 4 0 B -0 ~ 4 0 B -7 に接続されている。これらのポートバイパス回路は、選択信号 S E L(SEL-0A~SEL-7A、SEL-0B~SEL-7B)によって制御されている。

[0052]

一方、ドライブ基板BRD0は、マザーボード上のポートバイパス回路 7-0 Aの入出力端子 I NとOUTとの間に接続されたアクセスループ 4 0 A - 0 と、ポートバイパス回路 7-0 Bの入出力端子 I NとOUTとの間に接続されたアクセスループ 4 0 B - 0 を有し、ドライブ基板BRD0に搭載されたディスクドライブPDEV $0\sim$ PDEV3は、それぞれ 1 対のポートバイパス回路(7 0 - 0 A、7 0 - 0 B) \sim (7 0 - 0 A、7 0 - 0 B)を備え、ポートバイパス回路 7 0 0 A \sim 7 0 - 0 S

Aはアクセスループ40A-0に接続され、ポートバイパス回路70-0B~7 0-3Bはアクセスループ40B-0に接続されている。

[0053]

ドライブ基板BRD0上のポートバイパス回路(70-0A、70-0B)~(70-3A、70-3B)のSELECT線は、それぞれドライブPDEV0~PDEV3内のポートバイパス制御線と接続され、何れかのドライブが休止状態にある時、該ドライブがアクセスループ40A-0、40B-0からバイパスされるようになっている。

他のドライブ基板BRD1~BRD7上のディスクドライブも、上記ドライブ基板BRD0 と同様の構成で、基板上の1対のアクセスループ40A-j、40B-j (j=1~7)に接続されている。

[0054]

上記構成から明らかなように、第0群に属したディスクドライブPDEV0~PDEV3 1は、ドライブ基板およびマザーボードに形成されたディスクチャネル回線(アクセスループ)40A、40Bによって、互いに直列に接続されており、ディスクドライブに障害が発生した時、障害基板との接続ポートとなるポートバイパス回路の選択信号を"0"にすることによって、障害のあるドライブ基板をディスクチャネル回線40A、40Bから電気的に分離(バイパス)できるようになっている。

[0055]

第10図は、本発明のディスクアレイ装置において、ドライブ基板と該ドライブ基板上に搭載されたディスクドライブとの対応関係を示すために参照される基板管理テーブル80の1例を示す。

基板管理テーブル80は、ドライブ基板の識別番号(BRD)801をもつ複数のデータエントリ800-0~800-31からなる。各データエントリは、ドライブ基板識別番号(BRD)801と、図4に示したディスクアレイ構成における基板位置を示すX座標値802およびY座標値803と、基板の状態コード804と、搭載されているディスクドライブ(PDEV)の番号805との関係を示している。

[0056]

状態コード804の値iは、図7で示した搭載ディスクドライブの状態遷移S Ti(i=0~3)に応じて書き換えられる。従って、ディスクアレイに障害ドライブが1つも存在しない場合は、図10に示すように、状態コード804は、現用系として正常に動作中(状態ST0)であることを示す「0」、または予備ボード(状態ST1)であることを示す「1」の何れかの値となっている。

[0057]

第11図は、本発明のディスクアレイ装置において、障害回復処理のために参照されるドライブ管理テーブル81の1例を示す。

ドライブ管理テーブル81は、ディスクドライブ識別番号(PDEV)811 をもつ複数のデータエントリ810-0~810-127からなる。各データエントリは、ディスクドライブ識別番号(PDEV)811と、ディスクドライブが搭載されているドライブ基板の番号812と、ドライブ基板上でのディスクドライブの搭載位置を示すZ座標値813と、ディスクドライブのメモリ領域の最大論理ブロックアドレス814と、ディスクドライブの状態コード815との関係を示している。状態コード815の値iも、図7で示した状態遷移STi(i=0~3)に応じて書き換えられる。

[0058]

第12図は、本発明のディスクアレイ装置において、RAIDグループの構成 要素を定義したRAIDグループ管理テーブル82の1例を示す。

RAIDグループ管理テーブル82は、RAIDグループの識別番号(VDE V)821をもつ複数のデータエントリ820-0、820-1、・・・からなる。各データエントリは、RAIDグループ識別番号(VDE V)821と、マスタディスクアダプタ識別番号822と、RAIDタイプ823と、RAIDグループの構成ドライブを示すコンポーネントPDE V824と、RAIDグループが再編成された時、新たなRAIDグループの構成ドライブを示す新コンポーネントPDE V825との関係を示している。

[0059]

ここで、マスタディスクアダプタ識別番号822は、図1に示すように、ディ

スクアレイ 5 に 2 つのディスクコントローラ 1 A、 1 Bを接続したシステム構成において、マスタとなるディスクコントローラ側のディスクアダプタ 2 0 を示している。

[0060]

マスタディスクアダプタ識別番号822が「0」に設定されたRAIDグループは、ディスクコントローラ1A側のディスクアダプタ20がマスタとなり、「1」に設定されたRAIDグループは、ディスクコントローラ1B側のディスクアダプタ20がマスタとなる。マスタに指定されたディスクアダプタ20に障害が発生した場合は、他方のディスクアダプタ20が、障害ディスクアダプタに代わってディスクアレイ5をアクセスする。

[0061]

RAIDタイプ823は、RAIDグループに対して適用されるRAIDのタイプを示す。図示した例では、RAIDグループ $VDEV0\sim VDEV8$ にはRAID5が適用され、VDEV14、VDEV15にはRAID1が適用されていることを示している。

[0062]

以下、本発明のディスクアレイ装置におけるドライブ異常の回復処理について説明する。

図13は、現用系のディスクドライブに異常が検出された時、ディスクアダプタ20のプロセッサ24が実行する障害回復処理ルーチン200の1実施例を示す。

[0063]

図10~図12に示した管理テーブル80~82は、プロセッサ24がアクセスするローカルメモリ25内に形成され、障害回復処理ルーチン200の実行中に参照される。但し、これらのテーブルはキャッシュメモリ30に格納されてもよい。また、これらの管理テーブル80~82の内容は、SVP2を介して保守員に監視され、新エントリの設定とエントリ内容の変更が随時に行なわれる。

[0064]

プロセッサ24は、例えば、所定回数のリトライを繰り返してもデータが読み 出せない等、現用系のディスクドライブに異常が検出された時、SVP2に異常 発生を通知した後、障害回復処理ルーチン200を実行する。

プロセッサ 2 4 は、先ず、READ/WRITEアクセス中に障害が発生した異常ドライブPDEV(E)の識別番号 j に基づいて、ドライブ管理テーブル 8 1 から上記異常ドライブと対応するデータエントリ 8 1 0 ー j を検索し、該データエントリ 8 1 0 ー j の状態コード 8 1 5 の値を異常状態(ST2)を示す「2」に変更する。また、上記データエントリ 8 1 0 ー j から、異常ドライブが搭載されている障害基板BRD(E)の識別番号 8 1 2 の値 k を特定する(ステップ 2 1 1)。

[0065]

プロセッサ24は、上記障害基板BRD(E)の識別番号 k に従って基板管理テーブル80を参照し、データエントリ800-k から、障害基板BRD(E)の X 座標802の値Ex、Y 座標803の値Eyと、ドライブ番号欄805の内容から連累ドライブ番号PDEV(EZ1)~PDEV(EZ3)を特定する(ステップ212)。次に、基板管理テーブル80から、Y 座標803の値がEyで状態コード804が予備状態値「1」となっているデータエントリ800-yを検索し、障害基板BRD(E)に代わるべき交替基板BRD(R)の識別番号 y と、交替ドライブPDEV(RZ0)~PDEV(RZ3)の識別番号を特定する(ステップ213)。また、R A I D グループ管理テーブル82から、コンポーネントPDEV824に上記異常ドライブPDEV(E)の識別番号を含むデータエントリと、ステップ212で特定された連累ドライブ番号PDEV(EZ1)~PDEV(EZ3)の識別番号を含むデータエントリと、ステップ212で特定された連累ドライブ番号PDEV(EZ1)~PDEV(EZ3)の識別番号を含むデータエントリを検索する。これによって、再編成が必要となるグループ識別子VDEV(N1)~VDEV(N4)が特定される(ステップ214)。

[0066]

以下の説明では、再編成対象となったRAIDグループのRAIDタイプ823が全て「5」(レベル5)であったと仮定する。

プロセッサ 2 4 は、R A I D グループ管理テーブル 8 2 において、ステップ 2 1 4 で特定したグループ識別子VDEV (N1) ~ VDEV (N4) をもつデータエントリの新コンポーネントPDEV 8 2 5 として、上記異常ドライブPDEV (E) と連累ドライブ番号PDEV (EZ1) ~ PDEV (EZ3) の識別番号の代わりに、ステップ 2 1 3 で特定した交替ドライブPDEV (RZ0) ~ PDEV (RZ3) の識別番号を含む新たな組み合せ定義する(ステップ 2 1 5)。

[0067]

この後、プロセッサ24は、図14で詳述する交替ドライブPDEV(RZO)~PDEV(RZ3)のデータ再生/複写処理ルーチン220を実行し、障害基板BRD(E)で保持されていた全てのデータを交替基板上のディスクドライブに再現できた時点で、RAIDグループ管理テーブル82の新コンポーネントPDEV825の内容をコンポーネントPDEV824に書き写す(ステップ231)。上記コンポーネントPDEV824の内容変更によって新コンポーネントPDEVのデータが有効化され、図7で示したRAIDグループPDEVの再編成EVNT2が完了する。

[0068]

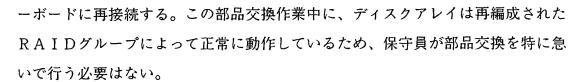
プロセッサ24は、PDEVの再編成の後処理として、ドライブ管理テーブル81において、交替ドライブPDEV(RZ0)~PDEV(RZ3)に対応するデータエントリの状態コード815の値を正常状態を示す「0」に変更し、異常ドライブPDEV(E)と連累ドライブ番号PDEV(EZ1)~PDEV(EZ3)に対応するデータエントリの状態コード815の値を交換待ち状態ST3を示す「3」に変更する(ステップ232)。また、基板管理テーブル80において、交替基板BRD(R)に対応するデータエントリの状態コード804の値を正常状態を示す「0」に変更し、障害基板BRD(E)に対応するデータエントリの状態コード804の値を交換待ち状態を示す「3」に変更する(ステップ233)。

[0069]

プロセッサ24は、上述したテーブル内容の更新が完了すると、障害基板BRD(E)の座標値Ex、Eyで特定される基板接続用のポートバイパス回路の選択信号を切替え、障害基板をディスクチャネル(アクセスループ)から切離し(ステップ234)、SVP2に対して、障害ドライブと障害基板を特定して、障害回復処理の完了と障害部品の交換を要求する制御メッセージを通知する(ステップ235)。

[0070]

上記制御メッセージの内容は、SVP2の表示装置9に出力される。従って、保守員は、上記制御メッセージに従って、障害ボードをマザーボードから取り外し、障害ドライブを正常ドライブに取り替え、部品交換後のドライブ基板をマザ



[0071]

保守員は、正常基板をマザーボードに再接続した後、表示装置9に付随する入力装置を利用して、ディスクアダプタ20のプロセッサ24に、障害基板の復旧を指令する。プロセッサ24は、上記指令に応答して、基板管理テーブル80の復旧基板と対応するエントリの状態コード804と、ドライブ管理テーブル81における上記復旧基板上の搭載ドライブと対応するエントリの状態コード815をそれぞれ交換待ち状態値「3」から予備状態値「1」に変更する。また、上記復旧基板BRD(E)の座標値Ex、Eyで特定されるポートバイパス回路の選択信号の状態を切替え、復旧基板上の搭載ドライブをディスクチャネルに接続する。

[0072]

図14は、データ再生/複写処理ルーチン220の詳細フローチャートの1例 を示す。

データ再生/複写処理ルーチン220では、先ず、異常ドライブPDEV(E)に格納されていたデータを交替基板BRD(R)上で上記異常ドライブと同一のZ座標値sをもつ交替ドライブPDEV(RZs)に再生するデータ再生処理300を実行する。データ再生処理300の詳細については、後で図16を参照して説明する。

[0073]

次に、連累ドライブの残り個数をカウントするためのパラメータiの値をクリア(ステップ221)した後、パラメータiの値をインクリメントする(ステップ222)。この後、障害基板BRD(E)上に搭載されている異常ドライブPDEV(E)以外のドライブPDEV(EZ1)~PDEV(EZ3)の中から選択された第i連累ドライブPDEV(EZi)の格納データをブロック単位で次々と読み出し、読み出されたデータを交替基板BRD(R)上で上記第i連累ドライブと同一のZ座標値iをもつ交替ドライブPDEV(RZi)に複写するデータ複写処理ルーチン400を実行する。データ複写処理400の詳細については、後で図17を参照して説明する。

[0074]

データ複写処理ルーチン400が完了すると、パラメータiの値が連累ドライブの基板搭載個数(この例では3個)に達したか否かを判定し(ステップ223)、連累ドライブの基板搭載個数に達していれば、このルーチン220を終了し、そうでなければ、ステップ222に戻って、次の連累ドライブについて同様の処理を繰り返す。

[0075]

図15は、データ再生処理300において、プロセッサ24が利用するデータ 再生管理テーブル83の構成を示す。

データ再生処理300では、図16で詳述するように、交替ドライブPDEV(RZs)への再生データの書込みが、複数データブロックからなる領域単位で繰り返される。

[0076]

データ再生管理テーブル83は、交替ドライブPDEV(RZs)のメモリ空間を複数のメモリ領域に分割して各メモリ領域に割当てた領域番号831と、メモリ領域の開始アドレスおよび最終アドレスを示す論理ブロックアドレス832と、再生済みフラグ833とを示す複数のエントリ830-1、830-1、…からなっている。再生データ書込み済みとなった領域には、再生済みフラグ833に値「1」が設定される。

[0077]

図16は、データ再生処理300の詳細フローチャートの1例を示す。

データ再生処理300では、プロセッサ24は、初期値として、パラメータSADDに交替ドライブPDEV(RZs)のメモリ空間における再生開始アドレスを設定し(ステップ301)、パラメータEADDに交替ドライブPDEV(RZs)のメモリ空間における再生終了アドレスを設定し(ステップ302)、パラメータBLに再生単位となる領域サイズ(またはデータブロック数)を設定し(ステップ303)、データ再生領域を切替えるためのパラメータCNTにパラメータSADDの値を設定する(ステップ304)。パラメータBLが示す領域サイズは、データ生成管理テーブル83における各論理ブロックアドレス83が示すアドレス範囲に相当している。

[0078]

次に、プロセッサ24は、パラメータCNTとEADDの値を比較する(ステップ305)。CNTの値がEADD以上の場合は、このルーチンを終了する。CNT<EADDの場合は、異常ドライブPDEV(Eのメモリ空間に対して上位装置からアクセス要求(READ/WRITE命令)があったか否かを判定する(ステップ306)。もし、アクセス要求がなければ、データ生成管理テーブル83を参照し、パラメータCNTで特定されるメモリ領域の再生済みフラグ833をチェックする(ステップ307)。

[0079]

再生済みフラグ833が「1」となっていた場合、プロセッサ24は、パラメータCNTにブロックサイズBLの値を加算し(ステップ310)、ステップ305に戻る。再生済みフラグ833が「0」となっていた場合は、既にRAIDグループ管理テーブル82のコンポーネントPDEV824で特定済みとなっている上記 異常ドライブPDEV(E)が所属するRAIDグループの他の正常ディスクドライブ から、パラメータCNTとBLで特定される1メモリ領域分のデータを読み出し、読み出されたデータのビット毎の排他論理和(ExOR)演算結果を交替ドライブ PDEV(RZs)の該当メモリ領域に次々と書き込む(ステップ308)。

この後、データ再生管理テーブル83のパラメータCNTで特定されるメモリ領域、すなわち、論理ブロックアドレス832がCNT~CNT+BLのデータエントリの再生済みフラグ833に「1」を設定し(ステップ309)、ステップ310を実行する。

[0080]

判定ステップ307で上位装置からアクセス要求を受けていた場合は、データ再生管理テーブル83を参照し、アクセス要求が示すアドレス範囲がデータ再生済みのメモリ領域に該当しているか否かを判定する(ステップ320)。アクセスすべきメモリ領域が、交替ドライブにデータ再生済みのメモリ領域に該当していた場合は、交替ドライブPDEV(RZs)を含む新RAIDグループに対してREAD/WRITE命令を実行し(ステップ323)、その後でステップ306を実行する。

[0081]

アクセスすべきメモリ領域が、未だデータ再生されていないメモリ領域に該当 していた場合は、RAIDグループの他の正常ディスクドライブから、アクセス 対象領域のデータを読み出し、読み出されたデータのビット毎の排他論理和(E \times O R)演算結果を交替ドライブPDEV(RZs)の該当メモリ領域に次々と書き込む(ステップ321)。この後、データ再生管理テーブル83で上記メモリ領域に該当するデータエントリの再生済みフラグ833に「1」を設定し(ステップ322)、ステップ323で上記交替ドライブPDEV(RZs)を含む新RAIDグループに対してREAD/WRITE命令を実行する。

[0082]

図17は、交替ドライブへのデータ複写処理400の詳細フローチャートの1 例を示す。

データ複写処理 400 において、プロセッサ 24 は、ステップ 401 ~ 404 で、データ再生処理 300 のステップ 301 ~ 304 と同様の初期値設定を行った後、パラメータ CNT と EADDの値を比較する(ステップ 405)。CNTの値が EADD 以上となっていた場合は、このルーチンを終了し、CNT 〈 EADD の場合は、現在、処理対象となっている連累ドライブ PDEV (EZi) のメモリ空間に対して上位装置からアクセス要求(READ/WRITE命令)があったか否かを判定する(ステップ 406)。

[0083]

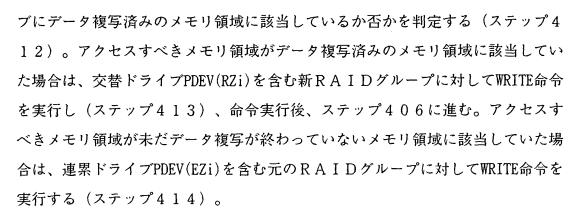
もし、アクセス要求がなければ、連累ドライブPDEV(EZi)から、パラメータCNT とBLで特定されるメモリ領域から複数データブロック分のデータを読み出し、これを交替ドライブPDEV(RZi)の該当メモリ領域に書き込む(ステップ4070)。この後、パラメータCNTにブロックサイズBLの値を加算し(ステップ4080)、ステップ4051に戻る。

[0084]

判定ステップ406で上位装置からアクセス要求を受けていた場合は、アクセス要求の種類を判定する(ステップ410)。アクセス要求がREAD命令の場合は、連累ドライブPDEV(EZi)のメモリ空間において上記READ命令を実行し(ステップ411)、命令実行後、ステップ406に進む。

[0085]

アクセス要求がWRITE命令の場合、アクセスすべきメモリ領域が、交替ドライ



[0086]

図13では、RAID5を前提に障害回復処理200の実施例を説明したが、 再編成対象となったRAIDグループのRAIDタイプ823が、例えば、RAID1(レベル1)の場合は、RAIDグループのコンポーネントが障害ドライブ と少なくとも1つの予備ドライブであり、データ再生/複写処理220では、予 備ドライブの格納データを交替ドライブに複写すれば済む。従って、コンポーネ ント数が少なくなるだけで、図13のステップ214~232と同様の手順で、 RAIDグループの再編成を実現できる。

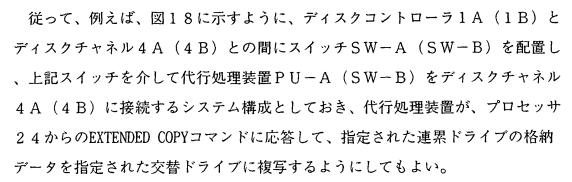
[0087]

以上の実施例では、ドライブ基板上の1つのディスクドライブに障害が発生した場合について説明したが、本発明は、同一基板上の複数のディスクドライブに障害が発生した場合でも適用できる。例えば、同一基板上の2個のディスクドライブに障害が発生した場合は、各障害ドライブに対してデータ再生処理300を実行し、残りの2個の連累ドライブに対してデータ複写処理400を実行すればよい。

[0088]

実施例では、データ再生/複写処理220をディスクアダプタ20のプロセッサ24で実行したが、データ再生/複写処理220の一部をプロセッサ24からの指令によって、他の処理装置に代行させるようにしてもよい。特に、連累ドライブから交替ドライブへのデータの複写は、SCSIにおけるEXTENDED COPYコマンドを利用して実行できる。

[0089]



[0090]

【発明の効果】

以上の実施例から明らかなように、本発明によれば、RAIDグループの再編成が完了した後に障害基板の部品を交換するようにしているため、障害基板の取り外し期間中でも通常のデータ・リード/ライトが可能となる。従って、本発明によれば、ドライブ基板の実装密度を高くし、且つ、障害ドライブの部品交換に十分な作業時間を許容できる。また、部品交換作業中に不在ドライブに対する論理グループによるデータ再生を行う必要がないため、上位装置からのディスクアクセス要求に対して迅速の応答することが可能となる。

【図面の簡単な説明】

【図1】

本発明が適用されるディスクアレイ装置の1実施例を示す図。

【図2】

チャネルアダプタ10の構成を示すブロック図。

図3】

ディスクアダプタ20の構成を示すブロック図。

【図4】

本発明によるディスクアレイ5の実装構造の1例を示す斜視図。

【図5】

ディスクドライブに異常が発生した場合に本発明で再編成すべきRAIDグループと交替基板との関係を示す図。

【図6】

再編成後のRAIDグループの構成を示す図。



本発明のディスクアレイにおけるディスクドライブの状態遷移を説明するため の図。

【図8】

ポートバイパス回路の構成図。

【図9】

ディスクドライブ群とディスクチャネルとの接続関係を示す図。

【図10】

障害回復処理で参照される基板管理テーブル80の構成図。

【図11】

障害回復処理で参照されるドライブ管理テーブル81の構成図。

【図12】

障害回復処理で参照されるRAIDグループ管理テーブル82の構成図。

【図13】

本発明による障害回復処理ルーチン200の1実施例を示すフローチャート。

図14]

図13の障害回復処理ルーチン200に含まれるデータ再生/複写処理ルーチン220の詳細フローチャート。

【図15】

データ再生/複写処理ルーチン220で参照されるデータ再生管理テーブルの 構成図。

【図16】

データ再生/複写処理ルーチン220におけるデータ再生処理300の詳細フローチャート。

【図17】

データ再生/複写処理ルーチン220におけるデータ複写処理400の詳細フローチャート。

【図18】

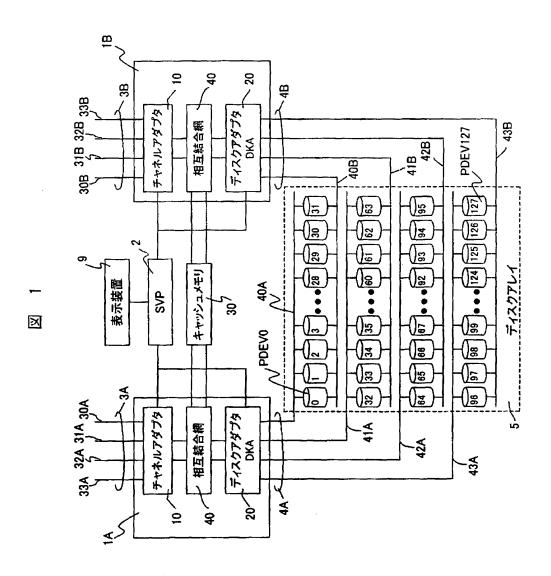
本発明が適用されるディスクアレイ装置の他の実施例を示す図。

【符号の説明】

- 1:ディスクコントローラ、2:SVP、3:チャネルパス、
- 4:ディスクチャネル、5:ディスクアレイ、6:共通基板(マザーボード)、
- 7:ポートバイパス回路、8:コネクタ、10:チャネルアダプタ、
- 20:ディスクアダプタ、11:ホストチャネルインタフェース、
- 12、21:キャッシュメモリインタフェース、
- 13、23:ネットワークインタフェース、14、24:プロセッサ、
- 15、25:ローカルメモリ、22:ディスクチャネルインタフェース、
- PDEV:ディスクドライブ、VDEV: RAIDグループ、BRD:ドライブ基板、
- SW-A、SW-B:スイッチ、PU-A、PU-B:代行処理装置。

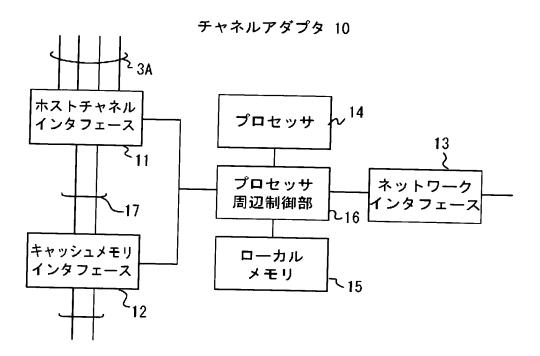
【書類名】図面

【図1】



【図2】

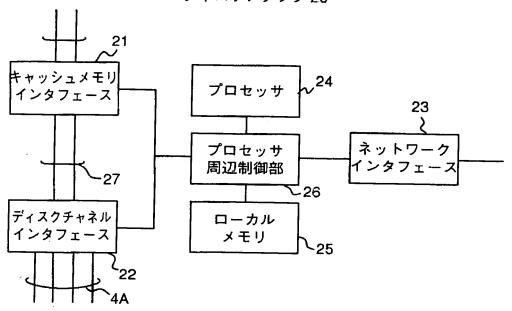
図 2



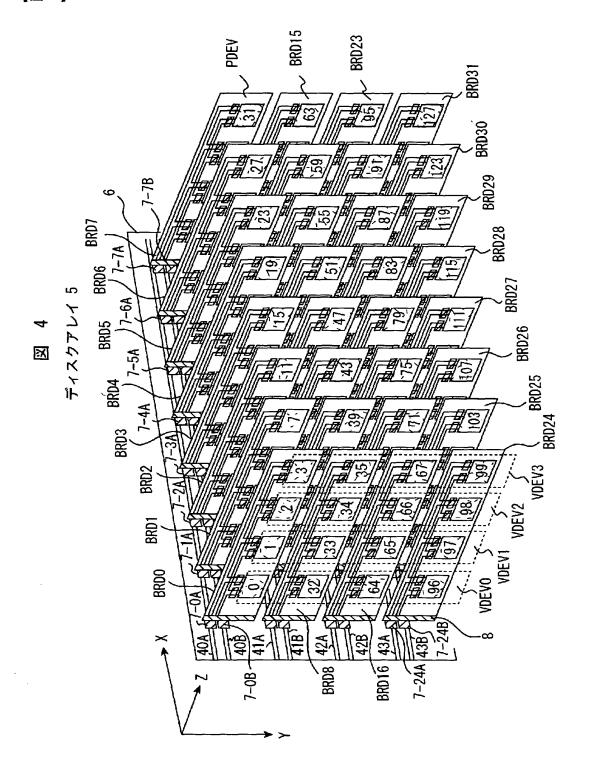
【図3】

図 3

ディスクアダプタ 20

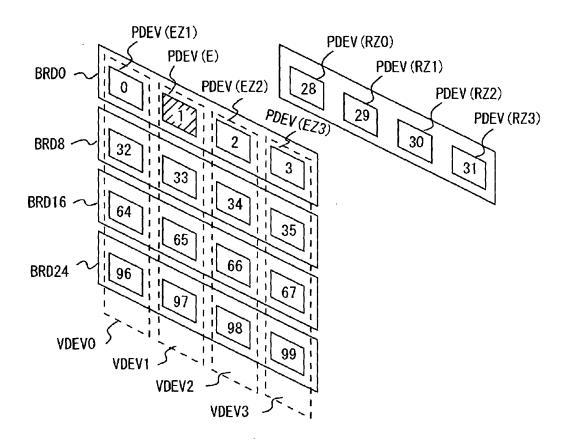


【図4】



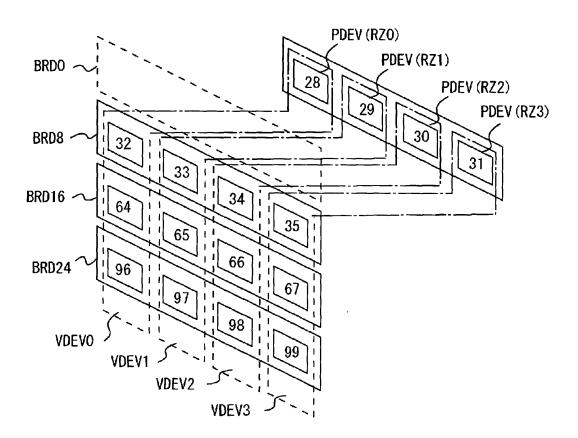
【図5】

図 5

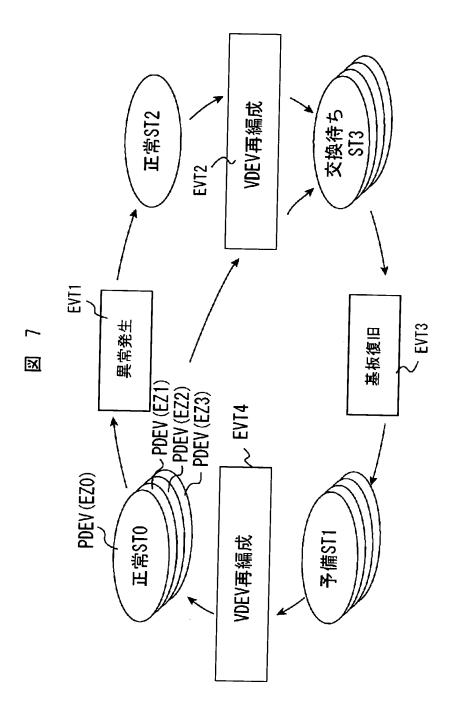


【図6】

図 6



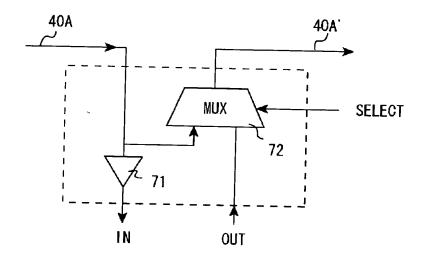
【図7】



【図8】

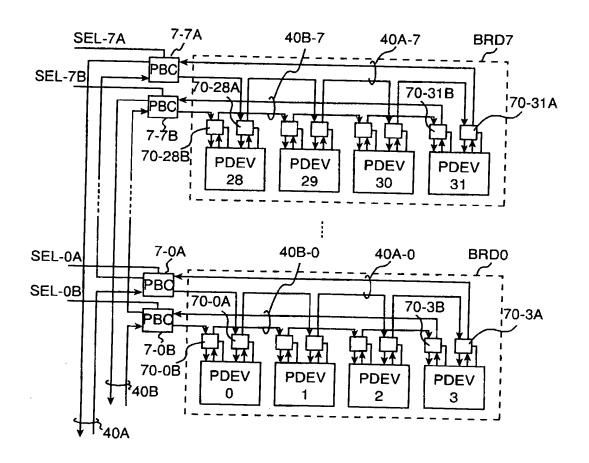
図 8

ポートバイパス回路 (PBC) 7



【図9】

図 9



【図10】

図 10

基板管理テーブル 80					
<u> </u>	§ 802	S 803	S 804	\$805	
基板番号 BRD	X座標	Y座標	基板状態 コード	ドライブ番号 PDEV	
0	0	0	0	$\begin{bmatrix} 0, & 1, & 2, & 3 \end{bmatrix} \sim 800-0$	
1	1	0	0	$4. 5. 6. 71 \sim 800-1$	
2	2	0	0	800-2	
3	3	0	0	[12, 13, 14, 15]	
4	4	0	0	{ 16, 17, 18, 19}	
5	5	0	0	[20, 21, 22, 23]	
6	6	0	1	{ 24, 25, 26, 27}	
7	7	0	1	[28, 29, 30, 31]	
8	0	1	0	[32, 33, 34, 35]	
9	1	1	0	36, 37, 38, 39	
10	2	1	0	40, 41, 42, 43	
11	3	1	0	{ 44, 45, 46, 47}	
12	4	1	0	48, 49, 50, 51	
13	5	1	0	52 , 53 , 54 , 55 }	
14	6	1	1	{ 56, 57, 58, 59}	
15	7	1	1	{ 60, 61, 62, 63}	
16	0	2	0	64, 65, 66, 67	
17	1	2	0	68, 69, 70, 71}	
18	2	2	0	[72, 73, 74, 75]	
19	3	2	0	176, 77, 78, 79	
20	4	2	0	80, 81, 82, 83	
21	5	2	0	84, 85, 86, 87	
22	6	2	1	88, 89, 90, 91	
23	7	2	1	92, 93, 94, 95	
24	0	3	0	96, 97, 98, 99	
25	1	3	0	{100, 101, 102, 103}	
26	2	3	0	{104, 105, 106, 107}	
27	3	3	0	{108, 109, 110, 111}	
28	4	3	0	[112, 113, 114, 115]	
29	5	3	0	{116, 117, 118, 119}	
30	6	3	1	[120, 121, 122, 123] 800-31	
31	7	3	1	{124, 125, 126, 127}	

【図11】

図 11

ドライブ管理テーブル 81					
<u></u>	∫ ⁸¹²	S 813	S 814	§ 815	_
PDEV	基板番号	Z座標	最大論理ブロックアドレス	PDEV 状態 コード	010.0
0	0	0	0x1FFFFFF	0 ′	~810-0
1	0	1	0x1FFFFFFF	0 -	~ 810-1
2	0	2	0x1FFFFFF	0	
3	0	3	0x1FFFFFF	0	
4	11	0	0x1FFFFFF	0	
5	11	11	0x1FFFFFF	0	
6	11	2	0x1FFFFFF	0	
7	11	3	0x1FFFFFF	0	
8	2	0	0x1FFFFFF	0	[
9	2	1	0x1FFFFFF	0	,
10	2	2	0x1FFFFFF	0	
11	2	3	0x1FFFFFF	0	
12	3	0	0x1FFFFFF	0	İ
13	3	11	0x1FFFFFF	0]
14	3	2	0x1FFFFFF	0	ļ
15	3	3	0x1FFFFFF	0	
16	4	0	0x1FFFFFF	0	
	:	:	:		
32	8	0	0x1FFFFFF	0	
33	8	1	0x1FFFFFF	0	·
34	8	2	0x1FFFFFF	0	
35	8	3	0x1FFFFFF	0	
36	9	1	0x1FFFFFF	0	
:	:	:	:		
122	30	2	0x1FFFFFF	1	
123	30	3	0x1FFFFFF	1	
124	31	0	0x1FFFFFF	1	
125	31	1	0x1FFFFFF	. 1	
126	31	2	0x1FFFFFF	1	810-127
127	31	3	0x1FFFFFF	1 -	V 121

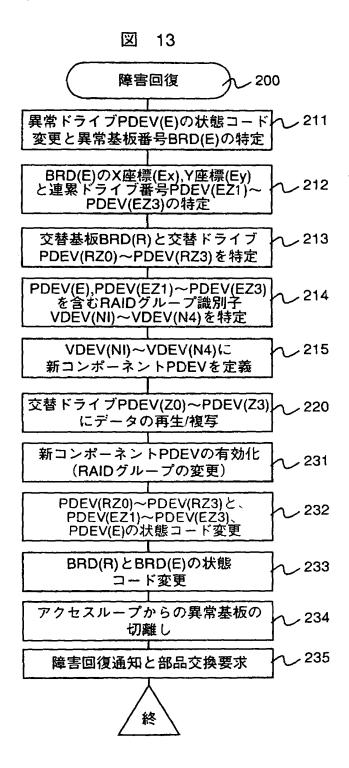
【図12】

図 12

RAIDグループ管理テーブル 82

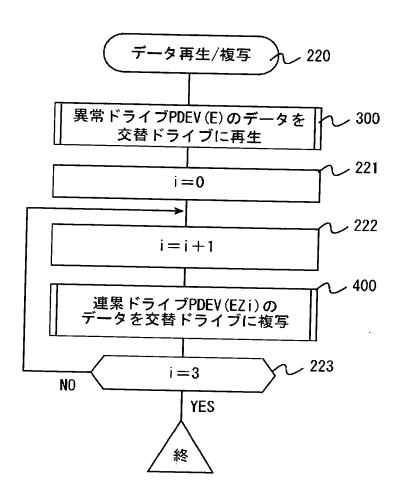
§ 821	S 822	823	§ 824	§ 825	
グループ 識別子 VDEV	マスタディスク アダプタ番号	RAID タイプ	コンポーネント PDEV	新コンポーネント PDEV	
0	0	5	{ 0, 32, 64, 96}	{28, 32, 64, 96}	820-1
1	0	5	1, 33, 65, 97	{29, 33, 65, 97}	~ 820-2
2	0	5	{ 2, 34, 66, 98}	{30, 34, 66, 98}] :
3	0	5	3, 35, 67, 99	(31, 35, 67, 99)	ļ
4	1	5	[4, 36, 68, 100]		}
5	11	5	[5, 37, 69, 101]		
6	1	5	6, 38, 70, 102		
7	1	5	{ 7, 39, 71, 103}]
8	0	5	8, 40, 72, 104		
:	:	:	:		
14	0	1	{14, 46}]
15	0	1	15, 47		}
:	:	:	:		

【図13】



【図14】

図 14



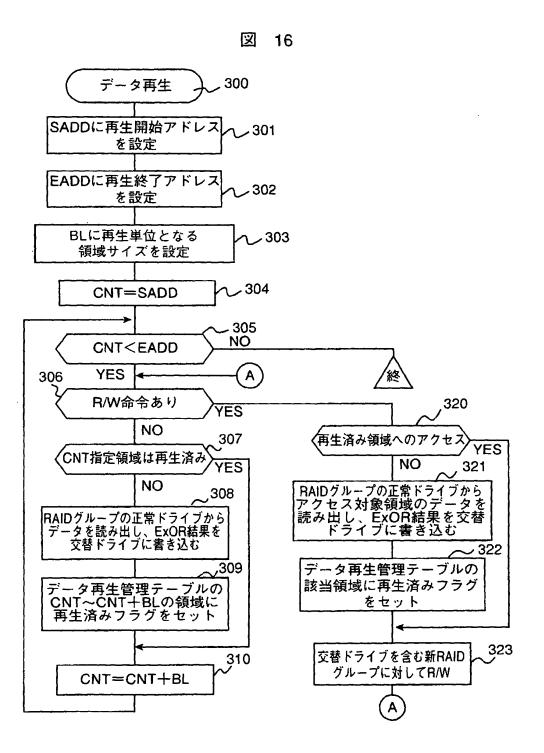
【図15】

図 15

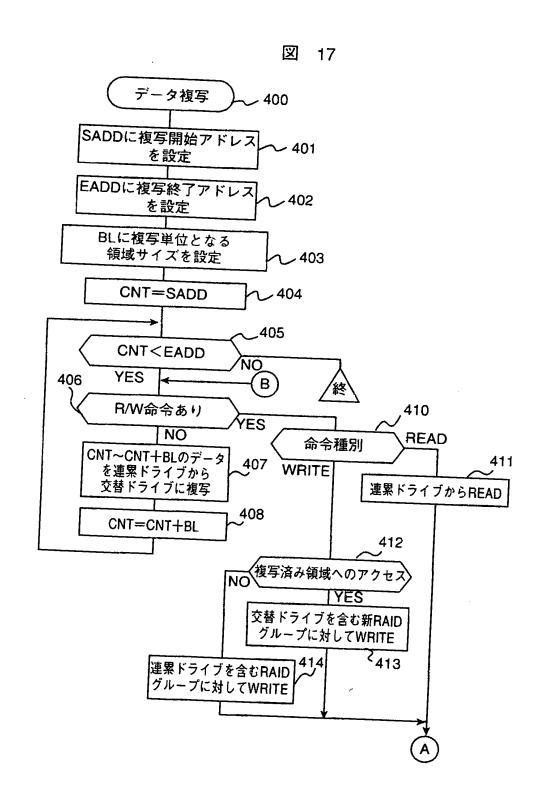
データ再生管理テーブル 83

831	832		833	
領域番号	論理ブロッ	再生済み	7	
	From	То	フラグ	
0	0×00000000	0x0000FFFF	1 ,	830-0
1	0x00010000	0x0001FFFF	1 .	830-1
2	0x00020000	0x0002FFFF	1 .	~830-2
3	0x00030000	0x0003FFFF	1	!
4	0x00040000	0x0004FFFF	0	
5	0×00050000	0x0005FFFF	0	
6	0×00060000	0x0006FFFF	0	
7	0×00070000	0x0007FFF	0	
8	0x00080000	0x0008FFFF	0	
9	0x00090000	0x0009FFFF	1	
10	0×000A0000	0x000AFFFF	0	



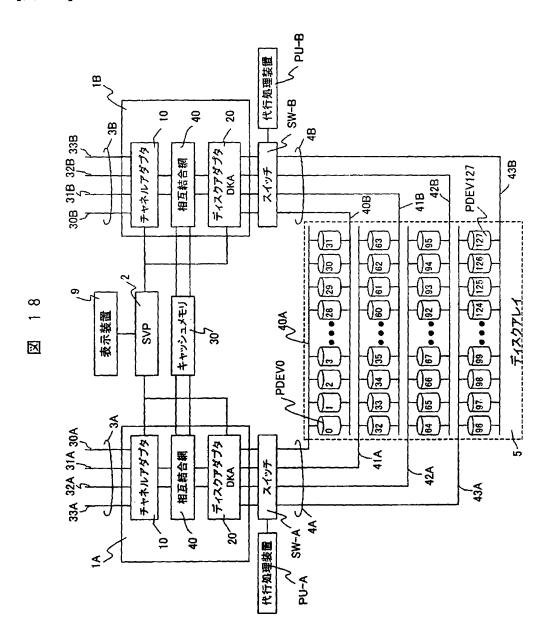








【図18】





【書類名】要約書

【要約】

【課題】 筐体サイズの小型化と記憶容量の大容量化に適し、ドライブ障害が発生しても、速やかに正常状態に復帰できるディスクアレイ装置および障害回復制御方法を提供する。

【解決手段】 上位装置と保守端末2に接続されたディスクコントローラ1と、ディスクチャネルを介して上記ディスクコントローラに接続されたディスクアレイ5とからなるディスクアレイ装置において、ディスクアレイにドライブ障害が発生した時、上記ディスクコントローラが、障害ドライブを搭載した障害基板上の複数のディスクドライブについて、交替ディスクドライブへのデータの移し替えを行い、論理グループの再編成が完了した後に、保守端末に障害基板が交換可能となったことを通知する。

【選択図】 図4



特願2003-061403

出願人履歴情報

識別番号

[000005108]

1. 変更年月日

1990年 8月31日

[変更理由]

新規登録

住 所 氏 名

東京都千代田区神田駿河台4丁目6番地

株式会社日立製作所